Statistical OT Lecture 2: Wasserstein Distances

1 More on Wasserstein Distances

To begin with we will investigate some properties of the Wasserstein distance and how it relates to other more familiar distances like the TV distance. One high-level takeaway is that Wasserstein distances are more sensitive to quantities like moments of the underlying distributions. We will discuss a series of useful facts that will help us build up this intuition. We will use the notation $\lesssim$ to denote an inequality which is true up to (usually universal) constants.

Fact: Suppose that μ, ν are supported on $[-1, 1]$ and define the vector of moments:

$$\alpha_j = \mathbb{E}_{X\sim\mu} X^j$$
$$\beta_j = \mathbb{E}_{Y\sim\nu} Y^j,$$

then we have the upper bound:

$$W_1(\mu, \nu) \lesssim \frac{1}{k} + 3^k \sum_{j=1}^{k} |\alpha_j - \beta_j|.$$

This fact is from a paper by Weihao Kong and Greg Valiant ("Spectrum Estimation From Samples"). Intuitively, this fact tells us that if two distributions have similar low-order moments then they are in fact (quantitatively) close in the W_1 sense. In contrast, the TV distance does not have such a relation to moments (imagine, taking a point mass and perturbing it slightly, this blows up the TV distance but keeps the moments stable).

Proof Sketch: The first thing we will need is the variational representation of W_1. We will derive this fact at some point in a future lecture. The W_1 distance has a dual representation as an "Integral Probability Metric" (IPM):

$$W_1(\mu, \nu) = \sup_{f\in \mathrm{Lip}_1} \mathbb{E}_\mu f - \mathbb{E}_\nu f,$$

where Lip_1 is the collection of 1-Lipschitz functions, i.e. functions which satisfy $|f(x) - f(y)| \leq L\|x - y\|$.

We will need to know some facts about the approximation of Lipschitz functions by polynomials. In particular, any Lipschitz function on $[-1, 1]$ can be approximated by a polynomial of degree k with uniform error (i.e. ℓ_∞ error) of order $1/k$, with coefficients $|c_j| \leq 3^k$. (You can find an explicit construction in the paper above or by looking up Jackson Theorems.)

With this fact in hand the proof is simple. For any Lipschitz function f, let us denote by $P_k(f)$ the polynomial whose construction we allude to above. Then we can write:

$$\begin{aligned}
\mathbb{E}_\mu f - \mathbb{E}_\nu f &= \mathbb{E}_\mu(f - P_k(f)) + \mathbb{E}_\nu(f - P_k(f)) + (\mathbb{E}_\mu - \mathbb{E}_\nu)(P_k(f)) \\
&\leq 2\|f - P_k(f)\|_\infty + \sum_{j=1}^{k} c_k(\alpha_j - \beta_j) \\
&\leq 2\|f - P_k(f)\|_\infty + \|c\|_\infty \sum_{j=1}^{k} |\alpha_j - \beta_j|,
\end{aligned}$$

which gives the claimed fact.

Now, we will show a sort of converse to the first result. Roughly, if the W_p distance between two distributions is small, then the distance between their moments is also (quantitatively) small. Concretely, we have the following fact:

Fact: Suppose that μ, ν are sub-exponential, then for any $\ell \geq 1$, $p > 1$ we have that:

$$|\mathbb{E}|X|^\ell - \mathbb{E}|Y|^\ell| \lesssim (p\ell/(p-1))^\ell W_p(\mu, \nu).$$

Background: It is useful to consult either Vershynin or Wainwright's books for some background on sub-exponential random variables. One definition (there are several equivalent ones) is that μ is sub-exponential if its MGF satisfies the bound:

$$\mathbb{E}\exp(|Z|) \leq 2.$$

Intuitively, sub-exponential random variables have very light tails. A direct application of Markov's inequality gives:

$$P(|Z| \geq t) \leq 2\exp(-t).$$

For a sub-exponential random variable we have the moment bound:

$$\mathbb{E}|Z|^k \leq 2k!$$

Proof: Let γ_0 denote the OT coupling of μ and ν for the ℓ_p cost. Using the fact that $x \to x^\ell$ is convex, we have by the mean value theorem:

$$\mathbb{E}_{(X,Y)\sim\gamma_0}\left[|X|^\ell - |Y|^\ell\right] \leq \ell\mathbb{E}_{\gamma_0}|X-Y|\max\{|X|^{\ell-1}, |Y|^{\ell-1}\}.$$

Applying Hölder's inequality we get:

$$\begin{aligned}\mathbb{E}_{(X,Y)\sim\gamma_0}\left[|X|^\ell - |Y|^\ell\right] &\leq \ell\left[\mathbb{E}_{\gamma_0}|X-Y|^p\right]^{1/p}\left(\mathbb{E}_{\gamma_0}\max\{|X|^{\ell-1}, |Y|^{\ell-1}\}^q\right)^{1/q} \\ &= \ell W_p(\mu,\nu)(\mathbb{E}_{\gamma_0}\max\{|X|^{\ell-1}, |Y|^{\ell-1}\}^q)^{1/q} \\ &\leq \ell W_p(\mu,\nu)(\mathbb{E}(|X|^{q(\ell-1)} + |Y|^{q(\ell-1)}))^{1/q}.\end{aligned}$$

Using the moment bound for sub-exponential random variables, we see that:

$$\begin{aligned}\mathbb{E}_{(X,Y)\sim\gamma_0}\left[|X|^\ell - |Y|^\ell\right] &\leq \ell W_p(\mu,\nu)(2q(\ell-1))^{(\ell-1)} \\ &\lesssim (q\ell)^\ell W_p(\mu,\nu).\end{aligned}$$

2 Relating Wasserstein distance to TV

In general, the Wasserstein distance can be much larger than the TV distance. For instance, take μ to be a point mass at 0, and ν to be an ϵ-contamination of μ which moves ϵ amount of the mass away to ∞. The Wasserstein distance will be ∞ but the TV distance is ϵ.

However, we can get around this by considering measures which are supported on a set of small diameter. Concretely:

Fact: Suppose μ, ν are supported on a set of diameter D, then:

$$W_p^p(\mu,\nu) \leq D^p \mathrm{TV}(\mu,\nu).$$

Proof: We will construct a coupling of μ, ν and use it to upper bound the Wasserstein distance. We will assume throughout that the measures μ, ν have densities f, g (we don't need the assumption but it makes things more transparent). Define the Yatracos set $A : f > g$.

Now, the TV distance has many equivalent representations. You can read more about this for instance in Chapter 2 of Tsybakov's book (Introduction to Non-Parametric Estimation). One of them is that:

$$\mathrm{TV}(\mu,\nu) = 1 - \int \min\{f,g\} := t.$$

Notice that, we can write:

$$f = \min\{f,g\} + \mathbb{I}_A(f-g)$$
$$g = \min\{f,g\} + \mathbb{I}_{A^c}(g-f).$$

We can then normalize each of these pieces to be a valid density to obtain that:

$$f = (1-t)\frac{\min\{f,g\}}{1-t} + t\frac{\mathbb{I}_A(f-g)}{t}$$
$$g = (1-t)\frac{\min\{f,g\}}{1-t} + t\frac{\mathbb{I}_{A^c}(g-f)}{t}.$$

This representation suggests a natural coupling of the two distributions (it is the one which "witnesses" the TV distance or alternatively is the optimal coupling for the TV distance). We let Z, Z_1, Z_2 be random variables with density $\frac{\min\{f,g\}}{1-t}$, $\frac{\mathbb{I}_A(f-g)}{t}$ and $\frac{\mathbb{I}_{A^c}(g-f)}{t}$ respectively. Then, we can obtain a coupling of μ,ν by the joint distribution of the pair $((1-B)Z + BZ_1, (1-B)Z + BZ_2)$, where B is a Bernoulli with parameter t.

Using this coupling it is straightforward to upper bound the Wasserstein distance:

$$W_p^p(\mu,\nu) \le P(B=1)\mathbb{E}\|Z_1 - Z_2\|^p \le D^p \mathrm{TV}(\mu,\nu).$$